Individual differences in teacher development

van der Lans, Rikkert M.; van de Grift, Wim J.C.M.; van Veen, Klaas

Published in:
Learning and Individual Differences

DOI:
10.1016/j.lindif.2017.07.007

Publication date:
2017

Individual differences in teacher development: An exploration of the applicability of a stage model to assess individual teachers

Rikkert M. van der Lans

Wim J.C.M. van de Grift

Klaas van Veen

Department of Teacher Education, Faculty of Social and Behavioral Sciences, University of Groningen, The Netherlands

Correspondence should be addressed to Rikkert van der Lans, Department of Teacher Education, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands. Telephone: +31 50 363 9754, E-mail: r.m.van.der.lans@rug.nl

Applicability of stage models to assess teacher development

Abstract

Researchers have recently become interested in exploring cumulative order in teachers' use of teaching practices, which they argue may reflect stages in teacher development. However, to validly apply stage models to individuals, it is necessary to determine whether all teachers fit the stage order. This study explores whether and in how many lessons observed teaching practices do not fit the stage order and whether misfit is typical to certain teachers, which would indicate individual differences. The sample consists of 198 classroom observations of 69 teachers (two to four lessons for each teacher). Using person-fit methods, the study shows that 17% of the 198 observed lessons substantially misfit the stage order but that misfit is not characteristic to specific teachers, suggesting that it is incidental. Removing the occasional misfitting lessons allows the stage model to provide an appropriate description of teaching skill.

Keywords: teacher development, teacher evaluation, stage model, person-fit statistics, Rasch model

1. Introduction

Scholars have recently advocated establishing stronger connections between research on teacher professional development and educational effectiveness research (EER) (Antoniou & Kyriakides, 2013; Authors, 2017; Kyriakides, Creemers, & Antoniou, 2009; Muijs et al., 2014). Traditionally, EER has addressed questions about what works in education (Muijs et al., 2014); within this tradition, teacher effectiveness research has generally focused on clusters of teaching practices associated with higher student achievement and learning (e.g., Brophy, 1986; Marzano, 2003; Muijs et al., 2014). This research stream has developed various observation instruments based on these findings (e.g., Kane et al., 2012; Strong, 2011) with the intention of informing teachers of how they perform in the classroom. However, consistent with the effectiveness tradition, in general classroom observation instruments focus on identifying effective teaching and do not address how it develops (e.g., Antoniou & Kyriakides, 2011, 2013; Authors, 2017). In response, both Author and Kyriakides have independently proposed stage models with the intention of finding the "developmental priorities of the teachers" (Antoniou & Kyriakides, 2013, p. 9) or, more specifically, tracing each teacher's "zone of proximal development" (Authors, 2017, p. 12). These stage models take a Vygotskian perspective (Palincsar, 1998; Vygotsky, 1978) on teacher development and argue that the success of teacher professional development depends on the match between feedback (and other learning materials) given to the teacher and his or her current development level.

However, although current findings suggest that stage models may provide an adequate description of the development of effective teaching practices for *most* teachers (Antoniou & Kyriakides, 2013; Authors, 2014, 2015, 2017; Kyriakides, Creemers, & Antoniou, 2009), evaluations and feedback have implications for individual teachers, and extant studies do not exclude the possibility that stage models provide an inadequate

description for a minority of teachers. Moreover, teacher development researchers have speculated that individual differences in teacher development are common (e.g., Berliner, 2001; Day, Sammons, Stobart, Kingston, & Gu, 2007; Sternberg & Horvath, 1995). If evaluators want to use stage models to advise individual teachers on directions for professionalization, training, and self-reflection, they must ensure that the particular teacher's development approximately fits with the stage model predictions. Therefore, the purpose of this study is to explore whether and how specific teachers' development aligns with the cumulative stages established by previous works (Authors, 2014, 2017).

2. Theoretical background

2.1 The International Comparative Analysis of Learning and Teaching stage model

The International Comparative Analysis of Learning and Teaching (ICALT) stage model identifies domains or stages of effective teaching practices. The term "effective teaching practice" refers to observable teaching practices, strategies, or methods that are positively related to students' achievement and school success, as described in, for example, Marzano (2003) and Muijs et al. (2014). Authors (2014) provides an extensive literature review elaborating how the stages are embedded in teacher effectiveness.

Two theories aid interpretation of the stages we observe: Fuller's (1969) stage theory of teacher development and Bloom, Engelhart, Furst, Hill, and Krathwohl's (1956) taxonomy of educational objectives. Fuller's (1969) theory first emphasizes the basic need for respectful relationships with students, which she refers to as the "self." Second, the theory identifies the need to acquire routines and procedures for classroom management and basic instructions ("tasks"). Fuller's third stage of teacher development focuses on teachers' need to improve their instructional practices and strategies ("impact"). To further refine Fuller's rather unspecific third stage, we turn to Bloom et al.'s (1956) taxonomy, which has been updated

and revised several times. We apply the terminology in Krathwohl's (2002) recent revision, which refers to Bloom's six categories of cognitive processing as remember, understand, apply, analyze, evaluate, and create. As Krathwohl elaborates, the revised taxonomy is hierarchical, reflecting stages in students' cognitive processing and learning. We adopt this perspective herein, maintaining that teachers' instructional practices can stimulate students to use these cognitive processes, and as such, they can be ordered cumulatively. We use the following six-stage model to describe teachers' skill development:

1. Learning to establish safe and respectful relationships. According to Fuller (1969), respectful relationships (herein also referred to as "climate") are among the first issues of classroom instruction that teachers must develop to become more effective. This critical role of respectful relationships is corroborated by psychological theory, including attachment (Bowlby, 1969) and self-determination (Ryan & Deci, 2000) theories. Attachment theory postulates that a safe environment stimulates children to take initiative and explore, because they know that an adult will be there to help them (Bowlby, 1969). According to Pianta and colleagues, the principles of attachment theory generalize to the classroom setting (Hamre et al., 2013; Pianta & Hamre, 2009). Wentzel's (2002) empirical findings suggest that students who view their teacher as fair and supportive (two key characteristics of our conceptualization of "respectful") are more likely to behave prosocially and thus are less likely to disturb classroom order and more likely to actively participate in academic activities. In addition, self-determination theory assigns a key role to respectful relationships in facilitating student motivation and performance (Ryan & Deci, 2000). Based on the above, our model predicts that respectful relationships are a requirement for orderly organized classrooms and successful instructions.

2. Learning to efficiently manage a classroom. Successful classroom management establishes procedures, routines, and rules about where and how learning takes place, which

are necessary for instructional activities to be executed successfully (Korpershoek, Harms, de Boer, van Kuijk, & Doolaard, 2016; Muijs & Reynolds, 2003). Teacher development theory generally assigns a key position to classroom management skills (Berliner, 2004; Fuller, 1969). If the classroom becomes disorganized, teachers typically focus on reestablishing adequate classroom management and postpone further instructional activities. If disorganization happens frequently, time to practice instructional skills becomes limited.

3. Developing clear and structured explanation skills. Clear explanation prompts students' prior knowledge, rehearses critical knowledge, and checks students' comprehension of the lesson content (Muijs & Reynolds, 2003; Rosenshine, 1995). Teacher development theory views explanations of assignments and tasks as part of management procedures, because these teaching practices tend to have a procedural character (e.g., Berliner, 2004; Fuller, 1969). Fuller (1969), for example, expects teachers' explanation skills to develop simultaneously with their skill in classroom management, suggesting that the two stages are indistinguishable in practice. However, the explanation domain is also the first in which teaching practices stimulate students to engage in cognitive processing of the lesson content. In terms of Bloom et al.'s (1956) taxonomy, clear explanation helps students remember and comprehend facts and procedures. Therefore, we distinguish explanation and management as two separate stages.

4. Developing skills in activating students. Successful activation stimulates interaction between teacher and students and among students—by, for example, collaborative group work, having students explain topics to one another, or having students think aloud (Abrami et al., 2015; Muijs & Reynolds, 2003). This stage and subsequent stages pertain to Fuller's third stage, "impact." Therefore, we apply Krathwohl's (2002) revision of Bloom et al.'s (1956) taxonomy to construct further understanding of what separates subsequent stages. In terms of Bloom et al.'s taxonomy, successfully activating students stimulates them to apply

and analyze the learned material. According to Bloom et al. and Krathwohl, students first need to remember and comprehend before they can apply this knowledge. Therefore, activating teaching practices can be successful only if the teacher has clearly explained the lesson content, which implies that teachers who lack routines to provide clear and structured explanation to students will have little time to deliberately practice how to activate students.

5. Learning to teach students learning strategies. Successful teaching of learning strategies enhances students' metacognitive skills and self-regulated learning—for example, by asking students to explain how they solved a problem or asking if there are multiple ways to answer the question (Abrami et al., 2015). In terms of Bloom et al.'s (1956) taxonomy, teaching learning strategies stimulates students to synthesize and evaluate the learned material. According to Bloom et al. and Krathwohl (2002), students first need to apply and analyze information before they can synthesize it with other knowledge or evaluate its value by taking different perspectives on the learned material. Thus, we maintain that teaching of learning strategies will be successful only if the teacher has successfully activated the student, which implies that teachers who have difficulty activating students will have little time to deliberately practice how to teach student metacognitive skills.

6. Developing skills in differentiation. Successful differentiation ensures that teachers adjust their instructional practice to specific students' learning needs by, for example, allowing flexibility in time to complete assignments or providing additional explanation to small groups (e.g., Reis, McCoach, Little, Muller, & Kaniskan, 2011). In terms of Bloom et al.'s (1956) taxonomy, differentiation involves helping low-ability students remember and comprehend, assisting moderate-ability students in applying and analyzing the material, and stimulating high-ability students in synthesizing and evaluating the material. Therefore, the model assumes that teachers must become skilled in all previous domains before they can truly differentiate. The word "truly" indicates that this logic allows for less sophisticated

differentiation. For example, teachers skilled in only the stages explanation and activation may differentiate between low-ability and moderate-/high-ability students. Thus, the theoretical proposition is that true differentiation is last in the ordering, but in observations of actual classroom practice, rudimentary differentiation may already be observed at stages 4 and 5.

Figure 1 illustrates the hierarchical and cumulative principle behind the model, in which skill in teaching practices of one stage is a prerequisite to developing skill in the next stage. Again, note that in practice, the boundaries between the stages are permeable: Although teachers learning to activate students have fewer opportunities to (deliberately) practice teaching students learning strategies than more advanced colleagues, they might occasionally find opportunities to do so.

---------- INSERT FIGURE 1 APPROXIMATELY HERE ----------

2.2 Similarities to other models of teacher development

The ICALT stage approach to teacher professional development exhibits similarities to the Dynamic Integrated Approach (DIA) to teacher professional development (Antoniou & Kyriakides, 2013; Kyriakides et al., 2009). For example, the DIA also cumulatively orders teaching practices and distinguishes five developmental stages (Antoniou & Kyriakides, 2013; Kyriakides et al., 2009) that reflect teaching effectiveness progression. Evidence that the cumulative order generalizes to other instruments strengthens the validity of the proposed cumulative approach. Furthermore, Antoniou and Kyriakides (2013) show that teachers receiving feedback and training based on the DIA cumulative stage model outperform teachers receiving feedback and training based on a holistic approach to professionalization. This finding provides some preliminary evidence of the stage models' underlying assumption

that feedback and training will be more effective if they match the teacher's current level of development.

In addition, other models of teacher development, including Berliner's (2001, 2004), Fuller's (1969), and Huberman's (1993), show similarities to ICALT's predictions. Previous research (Authors, 2015, 2017) argues that the six ICALT stages should mirror Fuller's (1969) three-stage theory of teacher development. Table 1 identifies similarities between Fuller's descriptions and the six stages presented herein. Fuller developed her theory in the context of student teachers and beginning teachers (Conway & Clark, 2003); although researchers currently view this stage theory as a valid description of the development of beginning teachers, they consider it too unspecific for evaluating more experienced teachers. Therefore, we add a comparison with Berliner's (2001, 2004) theory of development in teacher expertise, which more comprehensively describes experienced teacher development. Table 1 shows overlap between the current conceptualization and Berliner's (2004) five stages of teaching expertise: (1) novice, (2) advanced beginner, (3) competent, (4) proficient, and (5) expert. Berliner's model accentuates the cognitive differences in information processing, but he also suggests some examples of how these cognitive differences result in observable differences in teacher behavior.

---------- INCLUDE TABLE 1 APPROXIMATELY HERE ----------

2.3 Does one size fit all?

Literature on teacher development, including theories by Berliner (2001, 2004), Day et al. (2007), Fuller (1969) and Huberman (1993), routinely takes a Piagetian perspective (Piaget, 1964; Palincsar, 1998) on development, assuming that teachers' maturation or years of experience can determine their level of development. However, Berliner (2001) notes the

conceptual difficulties arising from this perspective, questioning the relationship between teachers' years of experience and growth in teaching skill. Evidence suggests that teaching skill increases with experience but that this increase quickly flattens (e.g., Rivkin, Hanushek, & Kain, 2005) and, in late career years, may even reverse (e.g., Day, 2008). This notion falsifies the assumption that years of experience is a good precursor of skill development.

Therefore, the here discussed stage models, like those of Author (2015, 2017) and Kyriakides et al. (2009), take a Vygotskian view on development (Palincsar, 1998; Vygotsky, 1978), which assumes that teachers' current or actual skill level is a better precursor of their skill development. In this view, teacher development is dependent on the match between actual and potential skill levels. Taking this perspective, this line of research has begun to explore the possibility of defining teachers' actual skill by stages of development (Authors, 2014, 2015, 2017; Kyriakides et al., 2009).

However, although evidence has confirmed the stages to be a good precursor of teachers' actual skill level, no research has explored the potential for individual differences in this definition. Thus, it is possible that for some specific teachers, the stages do not appropriately describe their actual skill level. To illustrate this, Figure 2 presents a hypothetical sample of four teachers. The actual skill level of the first three teachers is well defined by their stage, but that of the fourth teacher is not; that is, the fourth teacher shows skill in stages other than would be predicted according to his or her actual skill level.

---------- INSERT FIGURE 2 APPROXIMATELY HERE ----------

2.4 Research questions

The degree of rigidity of a stage theory is critical to its applicability. Several teacher development theorists have completely abandoned the idea of stages (e.g., Berliner, 2001;

Dall'Alba & Sandberg, 2006; Day et al., 2007; Huberman, 1993; Sternberg & Horvath, 1995), because they are unconvinced that the development of all teachers follows an identical sequence. This study therefore addresses two research questions. First, *in how many lessons are teachers' classroom practices inconsistent with the expected cumulative stage order?* To address this question, we examine the following hypotheses:

R1H$_0$: Teacher behavior in the observed lesson is consistent with the ICALT six-stage model of teacher development.

R1H$_1$: Teacher behavior in the observed lesson deviates from the ICALT six-stage model of teacher development.

Second, *is misfit typical to particular teachers?* This question explores whether misfit in stage order is repeatedly evident with the same teachers, thereby addressing the question of individual differences. To this end, we examine the following hypotheses:

R2H$_0$: Lessons in which teacher behavior misfits the ICALT six-stage model are uniformly distributed across teachers.

R2H$_1$: Lessons in which teacher behavior misfits the ICALT six-stage model are clustered within specific teachers.

3. Method

This research was embedded in a larger research project exploring how schools can organize and implement evaluation of in-service teachers using peer review and student ratings. The project was approved by the board of the teacher education department of the University of Groningen as being in accord with the principles and ethics of human subject research. School and teacher participation in the project was voluntary, and participating schools received no funding.

3.1 Participants

The sample consisted of 198 classroom observations of 69 teachers by 62 observers.[1] Most teachers taught Dutch (20%), English (as a foreign language; 20%), history (21%), and math (22%). The other 17% taught economy, geography, German, Latin, religion, science, social sciences, and technical drawing and construction. All teachers taught students in middle school (grades 7, 8, or 9). Teacher experience ranged from 1 to 40 years (M = 13 years, SD = 10 years), and 62.1% of the teachers were men. An analysis of variance (ANOVA) confirmed that the unrepresentative number of male teachers had few implications: The difference between men and women was negligible ($F(1, 196) = 1.756$, $p = .18$). The classroom observations took place from March 2014 through June 2014 and from February 2015 through June 2015. All observers also had teaching experience, which ranged from 1 to 40 years (M = 18 years, SD = 11 years), and 71.7% of the observers were men. A one-way ANOVA test verified that the overrepresentation of male observers had no effect on the evaluation results. The analysis confirmed no difference between male and female observers ($F(1, 196) = .01$, $p = .97$) or any indication of observer-gender × teacher-gender interactions ($F(1, 194) = .69$, $p = .56$).

3.2 Instrument

The ICALT is a Rasch-scaled observation instrument (Authors, 2014, 2017). The items refer to six domains or stages: safe learning climate, efficient classroom management, clarity of explanation (sometimes referred to as clarity of instruction), activating teaching methods, teaching learning strategies, and differentiation (see Table 2 for a sample of items). We used the complete instrument of Authors et al. (2017) and Authors et al. (2014), with the exception of item 22, “explains the lesson goals”; previous work consistently shows that it misfits the cumulative order (Authors, et al. 2014, Authors, et al. 2017).

[1] Authors et al. (2016) also uses this sample.

---------- INSERT TABLE 2 HERE ----------

Observers rated the items using four categories: 1 = "mostly weak," 2 = "more often weak than strong," 3 = "more often strong than weak," and 4 = "strong." However, when teachers are given feedback, the scores are dichotomized because the polytomous model does not add more information but presents additional complexity which teachers receiving the feedback find difficult to understand (Authors, 2017). The Pearson product correlation between evaluation outcomes of the polytomous partial credit model (PCM) and the dichotomous Rasch model is r = .92. The PCM evaluation outcomes ranged from –1.17 to 4.26 logits, which is smaller than that based on the Rasch model (–2.44 to 4.12 logits).

Studies have shown the instrument is predictive of student achievement test scores (Author et al. 1998) and student engagement (Author et al. 2014). Also, studies have shown that items of the "My Teacher" student survey relate to the same one-dimensional construct (Author et al. 2017).

3.3 Procedure and observation training

The research procedure was designed to simulate an actual implementation in schools. This procedure links to calls to study observation instruments inside actual schools (e.g., Cohen & Goldhaber, 2016; Peterson, 2000; Strong, 2011). Cohen and Goldhaber, for example, state that "much of what we know is derived from extensive research", but that it is unclear "how findings might translate when evaluation reform is put into practice" (p. 379).

To simulate an actual implementation several decisions were made in cooperation with the schools. First, schools have limited time and resources for observation training, so for this study, the training lasted four hours (as also is recommended by Strong, 2011), and observers must be considered "limitedly trained" in comparison to some previous studies (e.g., Kane et

al. 2012). In addition, we did not apply any tests or certification systems to prevent peer observers with insufficient interrater reliability from entering the classrooms. Most schools have limited or no access to statistics, such that an actual implementation would not involve the computation of interrater reliabilities (Strong, 2011). Furthermore, Peterson (2000) notes that schools are social organizations with their own group dynamics and they are unlikely to exclude willing peers from observing lessons. Therefore, the procedure excluded no one and all colleague-teachers were eligible to participate in the training regardless of their previous experience with classroom observation and regardless of their performances during the observation training. This research explores whether the proposed stage order can be used uniformly to evaluate individual teachers under these more realistic conditions.

The principal investigator provided observation training for all participants. The complete observation training took four hours. The training involved a half-hour introduction to the project and instrument, after which the trainees scored a lesson video and then participated in group discussion. We repeated the process with a second video. Lesson videos were specifically designed for the training and lasted 20 minutes each, but videos covered the start, middle and end of the lesson. The training also identified some ethical issues involved when visiting colleagues. Training took place inside the school, with groups no larger than 20 peers (most often approximately 8–12 peers).

We used four different videos for training, two per training session. We found the following interrater reliabilities for the four videos, calculated separately using a two-way random model and the polytomous response categories: Video 1: ICC = .88; 95% confidence interval (CI) = [.76, .95]; Video 2: ICC = .74; 95% CI = [.40, .94]; Video 3: ICC = .86; 95% CI = [.67, .97]; and Video 4: ICC = .88; 95% CI = [.79, .94]. In addition to assessing interrater reliability, we compared trainees' mean domain scores with sample mean scores provided by 188 trained observers. We then used this comparison to give trainees information about

whether they had been overly strict or lenient. We performed only one training and no follow-up, except when a school specifically requested it. After we gathered the data, the principal investigator of the project gave all participating teachers feedback in a 15- to 20-minute face-to-face conversation at the school.

3.4 Research design

We asked schools to group teachers in teams of four, and each team member visited one lesson of each teammate, which resulted in three classroom observations for each teacher. Lesson visits within the team involved the same class of students and spanned the complete lesson hour. Peers visited lessons alone, and the design involved no group visits. Observers were instructed to assign scores during the lesson visit. The time span between the first and last lesson visit was no longer than six months. The implemented design closely resemblances the design described by Ho and Kane's (2013, Table 10), except that ours has no administrator and all observations spanned the complete lesson hour. Some teachers received only two lesson visits because of situational circumstances; thus, the final data set consisted of 198 observations: 54 teachers (78%) received three lesson visits by three different colleagues, 14 teachers (20%) received two lesson visits by two colleagues, and two teachers received four lesson visits by three or four colleagues.

With regard to the observers, 18 colleagues (29%) visited one lesson, seven colleagues (11%) visited two lessons, 24 colleagues (39%) visited three lessons, and 14 colleagues (22%) visited more than three lessons. Despite our advice and encouragement, schools varied substantially in how they assigned observers. One school decided not to use colleague teachers; instead, three teacher coaches performed all 36 observations of nine teachers, which made it necessary to further examine differences between teacher-coaches and peer-colleagues. An independent *t*-test indicated no significant differences between evaluation scores of teacher-coaches and peer-colleagues ($M_{(\text{difference})} = .08$ [$SE = .24$], $t(df = 175) = .31$, p

= .75), suggesting that teacher-coaches did not evaluate teachers more leniently or strictly than peer-colleagues. Ho and Kane (2013) report that observations of peer-colleagues show less variation than formal administrators. Therefore, we decided to examine difference in the variation of observation scores. In this sample, teacher-coaches show slightly less variation in their evaluation scores (*SD* = 1.15) than peer-colleagues (*SD* = 1.34), suggesting that observations by peer-colleagues have more variation in performance between lessons than the more formal coaches. Consistent with the literature, teachers in general received favorable evaluations (Weisberg et al., 2009). That is, on average teachers received favorable ratings (i.e., a score of 3 or 4) on 23 of the 31 items.

3.5 Missing data

We instructed observers to score as many items as possible. If teaching practices were not observed, we asked observers to decide whether there were situations in which the teacher should have performed the behavior (in which case observers scored the item 1 = weak) or if there were no such situations (in which case observers scored the item missing). Of all item responses, 3% were reported missing; we regarded these missing values as missing at random.

3.6 Data analysis plan

To examine person misfit of the cumulative stage order, it is necessary to establish a baseline (i.e., cumulative order within the particular sample) and verify whether it reflects the six stages. To do so, we applied the Rasch model, specifically developed to estimate cumulative item order (Bond & Fox, 2007).

3.6.1 Estimating person fit. After the baseline order has been established, we evaluate whether each teacher fits the baseline order. For this end, the item response theory (IRT) person-fit statistic G_{NORMED} is used (Meijer, 1994). We note that G_{NORMED} is only one of multiple available person-fit statistics (for a recent overview of currently available person-fit statistics, see Tendeiro, Meijer, & Niessen, 2015). Researchers have developed person-fit

coefficients in general, and G_{NORMED} in particular, from Guttman (1944, 1950). According to Guttman, items have certain difficulties and individuals have certain skill, and valid measurement identifies the point at which the item difficulty matches the person's skill level. In Guttman's reasoning, an item may dominate the person (i.e., the item is too difficult for that person) or the person may dominate the item (i.e., the person is sufficiently skilled to perform well on the item). If items are ordered from highest to lowest scored, the pattern depicted in Table 3 should occur.

---------- INSERT TABLE 3 APPROXIMATELY HERE ----------

We define "error" as either when an item is estimated as too difficult but is performed well or when the person is estimated as sufficiently skilled but does not perform well on the item. For example, in Table 3, Teacher Y has a sum score of 3, from which the model would expect that this teacher had success with items 1, 2, and 3; however, Teacher Y did not have success with item 2 but had success with item 6. In general, the literature refers to these errors as "Guttman errors"; substantial discussion of how to count and weight Guttman errors has emerged, and various methods have been proposed (Mokken, 1971). Most are based on Guttman's (1950) reproducibility coefficient (Rep), which counts all erroneous responses and weights their number with the person's total score, though Mokken (1971) presents various psychometric arguments for why the Rep coefficient and its successors sometimes fail to correctly identify misfit and fit. It is beyond the scope of this article to review all proposed methods on how to count and weight Guttman errors (for a detailed discussion, see Mokken, 1971; Van Schuur, 2011). For our purposes, it is sufficient to note that Guttman's original proposal has become outdated and that researchers now regularly apply a definition of Guttman errors based on the "transivity relationship" between the person and an item-pair (for

details, see Van Schuur, 2011). This idea basically suggests that less weight be assigned to Guttman errors if errors are made with two items close in difficulty and that this weight should increase as the difference in item difficulty becomes greater.

The G_{NORMED} measure counts the number of Guttman errors for each individual teacher and divides this value by the total possible number. This calculation provides a coefficient that indicates the percentage of Guttman errors from 0.00 to 1.00. A G_{NORMED} of 1.00 indicates that the teacher shows higher-stage teaching practices and no lower-stage teaching practices (i.e., the observation is completely opposite to our predictions), .50 indicates that the teacher shows some lower- and some higher-stage teaching practices, and 0.00 indicates perfect fit (Table 3).

3.6.1 Defining the criterion. To identify teachers who do not fit the sequence, we must identify a criterion. No preset criterion exists for G_{NORMED}, so we must use another person-fit coefficient, specifically the chi-square test available in the eRm package, which uses the regular significance criterion $p < .05$ to identify misfit. The two person-fit tests appear to function similarly. Using these results, we identified the criterion of $G_{NORMED} > .30$.

3.7 Software

We estimated the baseline cumulative order using the R package eRm (Mair & Hatzinger, 2007), in which the "RM" argument estimates the item difficulties according to the dichotomous Rasch model and none of its default settings were changed. We estimated G_{NORMED} using the R package PerFit (Tendeiro et al., 2015), imputed missing values using the default nonparametric approach ("NPModel"), and estimated the cumulative order of the items using the one-parameter Rasch model. We then estimated the chi-square person-fit coefficient using the "personfit" argument available in eRm.

4. Results

Table 4 presents the overall cumulative ordering of teaching practices. Consistent with the stage model predictions, the order starts with the teaching practices of the lowest stage ("shows respect for student in behavior and language") and ends with those of the highest stage ("adapts processing of subject matter to student differences"). In general, most teachers dominate the teaching practices at the top of the table (i.e., most teachers use these practices effectively), whereas the teaching practices at the bottom of the table dominate most teachers (i.e., most teachers need to improve skill in stages preceding these teaching strategies). Certain practices cluster around similar *b*-values—suggesting that from a statistical standpoint, they are interchangeable—and at various points, the measurement scale shows gaps. The order in Table 4 is cumulative, and teachers showing practices located in the middle have most likely performed most of the practices above the middle, but not those below.

---------- INSERT TABLE 4 APPROXIMATELY HERE ----------

The order presented in Table 4 presumes that the items and people can be presented on a one-dimensional scale. Although previous studies provide evidence that observations of teaching practices fit with this assumption (Authors, 2014, 2017), for completeness we briefly reevaluated whether the assumption of one-dimensionality holds for our sample using Guttman's (1954) simplex factor analysis. To estimate this factor analysis, we used the CIRCUM program (Browne, 1992). The CIRCUM specifies an additional constraint that items have similar distances on the measurement scale, which, though overly strict for our purposes, cannot be removed. As Table 4 shows, the distance between *b*-parameters is not similar, and this constraint will therefore reduce model fit. We use the root mean square error of approximation (RMSEA) and view values below .05 as reflecting good model fit and values between.05 and.08 as reflecting modest model fit (Hu & Bentler, 1999). The results suggest modest model fit ($\chi^2(df = 431) = 920.27$, RMSEA = .076 [95% CI = .069, .083]),

which is sufficient for the study purposes (this study is not meant to further validate the instrument). Therefore, we made no further inspection of item (mis)fit.

4.1 Exploration of person misfit

This study aims to investigate whether and how many teachers' lessons show behaviors that misfit the hypothesized six stages of teacher development. For example, a lesson in which the teacher performs most behaviors associated with "clarity of explanation" (stage 3), while not performing many of the behaviors associated with "safe learning climate" (stage 1) or "efficient classroom management" (stage 2), misfits our predictions. Thus, we define misfit as substantially different cumulative ordering. We explore this notion using the person-fit statistic G_{NORMED}, and we also report results based on eRm-χ^2 person-fit statistic to validate our findings. A G_{NORMED} value of >.30 and a significant chi-square statistic (i.e., $\chi^2(df = 31) > 45.00$) indicate person misfit. Figure 3 displays the two person-fit statistics. The black line is G_{NORMED}, and the gray line is the chi-square statistic. The y-axis gives the number of classroom observations associated with that specific level of person fit, and the vertical line provides the cutoff values: The classroom observations on the left-hand side fit, and the classroom observations on the right-hand side misfit. As the figure shows, most classroom observations are consistent with the predicted stage order.

---------- INSERT FIGURE 3 APPROXIMATELY HERE ----------

Both $G_{NORMED} > .30$ and the chi-square test mostly identified the same lessons as deviating from the baseline cumulative order; however, if observation forms counted missing values, the two tests diagnosed some teachers differently, probably because of different imputation methods. Both criteria suggest that in approximately 17% of the lessons, the

observed teaching practices substantially deviate from those predicted by the teacher's actual skill level, which could reflect individual differences in development.

4.2 Is misfit typical to specific teachers?

To test for individual differences in development of teaching, we compared the fit of two nested random-effects logistic regression models using the chi-square difference test. The dependent variable is the nominal indicator 0 = misfit and 1 = fit. The first model specifies random effects for "school subject," "class," and "observer," and the second model adds the random effect "teacher." The rationale was to test whether "teacher" can account for clustering in misfit not already accounted for by the other context variables. Table 5 presents the results. The chi-square difference test rejects the alternative hypothesis and indicates that lessons in which observed teacher behavior misfit the stage order are approximately uniformly distributed across teachers ($\chi^2(df = 1) = .03$, $p = .87$.

---------- INSERT TABLE 5 APPROXIMATELY HERE ----------

Note that when we use a simple exact chi-square test, in which we do not include the other controlling variables, the results suggest individual differences in the stage ordering of effective teaching practices ($\chi^2(df = 69) = 93.60$, $p < .05$). Therefore, research implementing no control variables may report individual differences, however these differences may disappear when including the control variables.

5. Discussion and conclusion

Teacher evaluation research mainly concentrates on identifying ineffective teaching (e.g., Kane et al., 2012; Strong, 2011). However, identification alone is of little informative value to teachers. If evaluations are to make teachers more effective, they must identify not only (in)effectiveness but also teaching practices that the teacher can focus on improving next. Therefore, research behind the ICALT has directed attention to how to identify each

teacher's zone of proximal development. Although many questions around the ICALT stage model have yet to be answered, research addressing its validity, including this study, report positive results in general (e.g., Authors, 2014, 2015, 2017).

In taking this development stage model perspective, perhaps one of the most methodologically challenging aspects is how to explore individual differences. Previous studies validating the stage ordering (e.g., Authors, 2014, 2017) have largely focused on the sample average and provide no indication about whether the established stages apply to specific teachers. However, teacher evaluation and feedback typically affect individual teachers. To address this research gap, the current study explores the potential of person-fit methods for identifying specific lessons that misfit the stage model and finds that approximately 17% of the lessons do so.

However, we did not find that misfit clustered around specific teachers, which suggests that the observed misfit does not reflect individual differences between teachers. This finding is important because it provides a nuanced view of the concerns about whether stage models in general can provide an appropriate description of every teacher's development (e.g., Dall'Alba & Sandberg, 2006; Day et al., 2007; Huberman, 1993; Sternberg & Horvath, 1995). Providing an adequate description of each teacher's actual skill level seems possible with the ICALT stage model. Note that the reported findings pertain to individual differences in teachers' actual skill levels. The study interprets these findings in terms of teacher development because it assumes teachers' actual skill level to be the precondition of development. However, we acknowledge that measurement of teachers' actual skill level requires only one time point, and thus given the lack of multiple time points, some critics may view our interpretation of development as inadequate.

5.1 Possible effects of the observation procedure and observation training

The research procedure was designed to simulate what a real-world implementation of the classroom observation instrument in schools would involve. The consequence of working within the schools was that observer training was limited. In addition, although we shared information about interrater reliability and criterion validity with schools during the training, they required that we not use them as means to prevent unreliable observers from entering the classroom. This request raises questions whether our findings might have been affected by these lenient observation training standards. However, if the limited observation training and standards had effect, we would expect it to elevate the number of errors and misfitting lessons, such that the 17% reported herein would be an overestimation of what can be achieved when working with expert and/or extensively trained observers. In another dataset consisting of 567 lesson observations by extensively trained teacher educators we found only 6% of the lessons to misfit (Authors, 2017). This confirms the impression that additional training might further lower the number of errors and model misfit.

Another point of concern involves the procedure, in which teachers are also observers and thus know the exact content of the instrument. Some critics might argue that teachers could have manipulated their scorings by performing teaching practices that they otherwise would not. However, we observed no indications that teachers were (capable of) doing so. Such manipulation should result in greater homogeneity in scorings and ceiling effects, and a generalizability study indicated typical amounts of between-teacher and within-teacher variances (Authors, 2016). Moreover, the average score of 23 was well below the ceiling of 31 points and seems consistent with other observation instruments applied by schools (Weisberg et al. 2009).

5.2 Critical reflection on the operationalization of the safe learning climate domain

In Section 2.2, we compare the ICALT stage model with the DIA, which also uncovers a cumulative order that could possibly reflect development stages. However, the

DIA differs somewhat from ICALT, most profoundly in that the former presupposes that the quality of relationships and learning climate remains important in subsequent stages (Antoniou & Kyriakides, 2013; Kyriakides et al., 2009), while the latter concentrates on relationship and climate in the single stage "safe learning climate." This difference implies that the ICALT might have a narrower conceptualization of the constructs "relationships" and "climate" than the DIA.

This impression is further corroborated when comparing the ICALT framework with other teaching effectiveness models, specifically the teaching through interaction (TTI) model (Hamre et al., 2013). At a basic level, the TTI and ICALT models share similarities and overlap. Although the TTI clusters teaching practices into three domains (emotional support, classroom management, and instruction) and the ICALT clusters them into six, the ICALT's final four domains all involve aspects of classroom instruction. Thus, ICALT may be interpreted as distinguishing "safe learning climate," "classroom management," and a variety of instructional practices (broadly instruction). However, like the DIA and unlike the ICALT, the TTI framework emphasizes the emotional support domain, further subdividing it into four dimensions. In summary, the ICALT framework may be further improved with application of a broader conceptualization of "relationship" and an exploration of the potential role of relationships and learning climate in subsequent stages. Further research is currently performed to explore these issues.

5.3 Implications for policy

The results indicate that in a sizable minority of the lessons (17%), teaching behaviors are inconsistent with the ICALT stages. Thus, if ICALT is applied to discern teachers' stage of development, schools should ensure that teachers are observed on more than one occasion to lower the chances of misidentifying certain teachers' actual skill level. In line with previous work (Authors, 2016; Kane et al., 2012), we recommend gathering multiple classroom

observations for each teacher to ensure that at least some of the observations fit the cumulative order.

5.4 Limitations

To identify lessons in which teachers' behavior substantially deviates from the predicted cumulative order, it is first necessary to define the word "substantially." This definition requires determining how many Guttman errors are acceptable, and the current results completely rely on the validity of this criterion. If the criterion is too lenient, no individual differences can ever be discerned; if the criterion is too strict, the statistic identifies even rather unimportant individual differences. This study used a preset criterion, specifically the significance test, to define $G_{NORMED} > .30$ as "substantial." We can also defend use of this value on logical grounds: The pattern cannot be random or reversed and thus will need to show correspondence to the baseline cumulative order. Thus, our criterion excludes extreme alternative developmental patterns but does not diagnose all the lesser extremes. Further research might show that this criterion was too lenient.

References

Authors (1998). Blinded for review

Authors (2014). Blinded for review

Authors (2014). Blinded for review

Authors (2015). Blinded for review

Authors (2016). Blinded for review

Authors (2017). Blinded for review

Authors (2017). Blinded for review

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, *85*(2), 275-314. doi: 10.3102/0034654314551063

Antoniou, P., & Kyriakides, L. (2011). The impact of a dynamic approach to professional development on teacher instruction and student learning: Results from an experimental study. *School Effectiveness and School Improvement, 22*(3), 291–311. doi: 10.1080/09243453.2011.577078

Antoniou, P., & Kyriakides, L. (2013). A dynamic integrated approach to teacher professional development: Impact and sustainability of the effect of improving teacher behavior and student outcomes. *Teaching and Teacher Education, 29,* 1–12. doi: 10.1016/j.tate.2012.08.001

Berliner, D. (2001). Learning about learning from expert teachers. *International Journal of Educational Research, 35,* 463–483.

Berliner, D. C. (2004). Expert teachers: Their characteristics, development and accomplishments. In R. Batllori i Obiols, A. E Gomez Martinez, M. Oller i Freixa, & J. Pages i Blanch (Eds.), *De la teoria….a l'aula: Formacio del professorat ensenyament de las ciències socials* (pp. 13–28). Barcelona, Spain: Departament de

Didàctica de la Llengua de la Literatura i de les Ciències Socials, Universitat Autònoma de Barcelona.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.) (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* New York: David McKay.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Bowlby, J. (1969). *Attachment and loss: Vol. 1. Attachment.* New York: Basic Books.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist, 41*(10), 1069–1077.

Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrics, 57,* 469–497.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378-387.

Conway, P. F., & Clark, C. M. (2003). The journey inward and outward: a re-examination of Fuller's concerns-based model of teacher development. *Teaching and Teacher Education, 19,* 465–482.

Dall'Alba, G., & Sandberg, J. (2006). Unveiling professional development: A critical review of stage models. *Review of Educational Research, 76*(3), 383–412.

Day, C. (2008). Committed for life? Variations in teachers' work, lives and effectiveness. *Journal of Educational Change, 9*(3), 243–260.

Day, C., Sammons, P., Stobart, G., Kingston, A., & Gu, Q. (2007). *Teachers matter: Connecting lives, work and effectiveness.* Maidenhead, UK: Open University Press.

Fuller, F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6,* 207–226.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, E. F. Lazarsfeld., S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (Vol. 4, pp. 60–90). Princeton, NJ: Princeton University Press

Guttman, L. L. (1954). A new approach to factor analysis: The radex. In P. F. Lazersfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258–348). Glencoe, IL: The Free Press

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teaching effectiveness in over 4,000 classrooms. *The Elementary School Journal, 113,* 461–487.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill & Melinda Gates Foundation.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1–55.

Huberman, M. (1993). *The lives of teachers.* New York: Teachers College Press.

Kane, T. J., Staiger, D. O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., Kerr, K., Kawakita, T., & Parker, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management

programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research, 86*(3), 643–680.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218.

Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behavior and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25,* 12–23.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modelling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20,* 1–20.

Marzano, R.J. (2003). *What works in schools: Translating research into action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311–314.

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research* (Vol. 1). The Hague, NL: Walter de Gruyter.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231–256, doi: 10.1080/09243453.2014.885451

Muijs, D., & Reynolds, D. (2003). Student background and teacher effects on achievement and attainment in mathematics: A longitudinal study. *Educational Research and Evaluation, 9*(3), 289–314. doi: 10.1076/edre.9.3.289.15571

Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of Psychology, 49*(1), 345–375.

Piaget, J. (1964). Part I: Cognitive development in children: Piaget development and learning. *Journal of research in science teaching, 2*(3), 176-186.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practice.* Thousand Oaks, CA: Corwin Press.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational researcher*, *38*(2), 109–119.

Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal*, *48*(2), 462–501.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rosenshine, B. (1995). Advances in research on instruction. *Journal of Educational Research*, *88*(5), 262–268.

Ryan, R., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78.

Sternberg, R. J., & Horvath, J. A. (1995). A prototype view of expert teaching. *Educational Researcher, 24*, 9–17.

Strong, M. (2011). *The highly qualified teacher: What is teacher quality and how do we measure it?* New York: Teachers College Press.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2015). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1–27

Van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis* (Vol. 169). Thousand Oaks, CA: Sage Publications.

Vygotsky L. 1978. *Mind in Society: The Development of Higher Psychological Processes,* eds., M. Cole, V. John-Steiner, S. Scribner, E. Souberman. Cambridge, Massachusetts: Harvard University Press

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York: New Teacher Project.

Wentzel, K. R. (2002). Are effective teachers like good parents? Teaching styles and student adjustment in early adolescence. *Child Development*, *73*(1), 287–301.

Table 1

Overlap between the ICALT six-stage model, Berliner's (2004) five-stage model, and Fuller's three-stage model of teacher development

ICALT stage		Berliner stage (Berliner, 2004, pp. 19–23)		Fuller stage (Conway & Clark, 2003)	
1. Safe learning climate	Teacher is learning how to apply rules and keep order, without becoming overly rigid and disrespectful to (individual) students. Effective time for instruction is minimal.	**1. Novice**	Shows minimal skill at the task of teaching, conforming to whatever rules and procedures they were told to follow and is relatively inflexible.	**1. Self**	Concerns with meeting others' expectations: "I fear they would not listen or think of me as a teacher." Concerns with relations and attitude: "I hope to be more relaxed and have fun in teaching" (p. 473).
2. Efficient classroom management	Teacher is learning how to organize classroom activities and changeovers between activities. Management is still too disordered to keep most students on task or attend to instructions.	**2. Advanced beginner**	Has difficulties with students challenging the teacher's authority; students neurotically seek the teacher's attention. Teachers are learning when to ignore or break the rules, how to praise students and give them feedback.	**2. Task**	Concerns with clarity of instructions, assignments, and materials: "I hope I can teach all subjects effectively", "I hope children can all have the same resources", "I hope I can learn more about teaching subject matter to kindergartners", and "I hope I can manage the kids—the whole class by myself" (p. 471).
3. Clarity of explanation	Teacher can apply rules without becoming too rigid or disrespectful and can organize classroom activities and changeovers between them. The teacher is learning how to organize frontal class instructions, how to give feedback and developing routines how and when to check whether students understand.				
4. Activating teaching methods	Teachers have adequate management skills and give clear instructions, but they struggle with how to cope with and manage the fast and less predictable nature of more interactive and collaborative teaching approaches.	**3. Competent**	Stops making timing errors, can identify which student can do a task or as the culprit in a classroom problem but are not yet very fast, fluid or flexible in their behavior.	**3. Impact**	Not reported by Conway and Clark (2003). Fuller (1969) reports that experienced teachers' concerns focus on pupil gain and self-evaluation as opposed to personal gain and evaluation by others. Some concerns explicitly reported are about ability to understand students' capacities, how to specify objectives for students, and how to assess and evaluate pupil gains.
5. Teaching students learning strategies	Teachers can manage interactive instruction, which gives them rich feedback about what students are doing and thinking. They can identify students' misconceptions and confusion, but still too often fall back on giving students the answer, instead of having them to find it themselves. The teacher only infrequently learns students' meta-cognitive skills.	**4. Proficient**	Can predict when students start to act out, when the class gets bored, or when students are confused or exited.		
6. Differentiation	The teacher has learned how to teach students meta-cognitive skills. Students can monitor their own learning. This gives the teacher time to learn how and when to adjust instructions for particular students if the situation or context requires. adaptation.	**5. Expert**	Do things that usually work. Their performance seems fluid. Can handle classroom situations in which anomalies occur or when something atypical is noted.		

Table 2

Example items and indicators operationalizing the six domains

Domain	**Example items**	**Example indicators**
Safe learning climate	Shows respect for the pupils in behavior and language use	- Allows pupils to finish speaking - Listens to what pupils say
Efficient classroom management	Uses the time for learning efficiently	- Starts the lesson on time - Does not keep pupils waiting
Clarity of explanation	Gives clear explanation of how to use didactic aids and how to carry out assignments	- Explains how lesson aims and assignments relate to each other - Explains clearly which materials and sources can be used
Activating teaching methods	Stimulates pupils to think about solutions	- Shows pupils they can take towards a solution - Shows learners how to consult sources and reference works
Teaching students learning strategies	Teaches pupils how to simplify complex problems	- Teaches pupils how to break down complex problems into simpler ones - Teaches pupils to order complex problems
Differentiation	Adjusts instructions to relevant interlearner differences	- Gives additional instructions to small groups or individual pupils - Does not simply focus on the average learner

Table 3
Example of five teacher-specific scoring patterns

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	**Valid**	**G**$_{\textbf{NORMED}}$
Teacher A	1						**Yes**	.00
Teacher B	1	1	1				**Yes**	.00
Teacher C	1	1	1	1	1		**Yes**	.00
Teacher …								
Teacher Y	1		1			1	**No**	.44
Teacher Z				1	1	1	**No**	–1.00

Notes. The six hypothetical items are ordered cumulatively from easy to complex. The score of 1 indicates observed (success), and an empty cell indicates not observed (failure).

Table 4
Ordering of effective teaching practices from least (top) to most (bottom) complex

Stage	Teaching practice	b	SE
Climate	Shows respect for students in behavior and language	–2.35	.373
Explanation	Explains the subject matter clearly	–1.96	.326
Climate	Creates a relaxed atmosphere	–1.68	.295
Climate	Ensures mutual respect	–1.43	.273
Climate	Supports student self-confidence	–1.40	.274
Explanation	Gives well-structured lessons	–1.40	.273
Management	Ensures the lesson runs smoothly	–1.34	.267
Management	Ensures effective class management	–1.21	.256
Management	Uses learning time efficiently	–1.01	.243
Explanation	Gives feedback to students	–.69	.225
Explanation	Encourages students to do their best	–.57	.217
Management	Checks during processing whether students are carrying out tasks properly	–.27	.209
Explanation	Involves all students in the lesson	–.27	.203
Explanation	Clearly, explains teaching tools and tasks	–.31	.211
Activation	Asks questions that encourage students to think	–.31	.204
Activation	Encourages students to reflect on solutions	–.18	.200
Activation	Uses teaching methods that activate students	–.11	.196
Activation	Has students think out loud	–.02	.194
Activation	Provides interactive instruction	.25	.187
Explanation	Checks during instruction whether students have understood the subject matter	.33	.183
Learning strategies	Encourages students to apply what they have learned	.47	.182
Activation	Boosts the self-confidence of weak students	.49	.181
Learning strategies	Teaches students how to simplify complex problems	.98	.178
Learning strategies	Encourages students to think critically	1.12	.174
Learning strategies	Encourages the use of checking activities	1.45	.176
Learning strategies	Teaches students to check solutions	1.35	.177
Learning strategies	Asks students to reflect on approach strategies	1.70	.178
Differentiation	Checks whether the lesson objectives have been achieved	1.72	.175
Differentiation	Adapts instruction to relevant student differences	2.21	.179
Differentiation	Offers weak students additional learning and instruction time	2.22	.180
Differentiation	Adapts processing of subject matter to student differences	2.23	.181

Table 5
Results of random-effect logistic regression (dependent variable 0 = misfit, 1 = fit)

	Model 1		Model 2	
Fixed effects	***b***	**95%CI**	***b***	**95%CI**
Intercept	2.04	1.35, 3.15	2.06	1.35, 3.15
Random effects	**σ^2**	**95%CI**	**σ^2**	**95%CI**
Class	1.42	.57, 2.24	1.42	.52, 2.25
Observer	.00	.00, 1.07	.00	.00, 1.07
Subject	.01	.00, 1.32	.00	.00, 1.32
Teacher			.09	.00, 1.46
Deviance	166.54		166.51	

Figure 1.

Staged progression of teacher development of effective teaching. Check boxes indicate that the teaching practices associated with this stage are observed, and crosses indicate that behaviors are not observed.

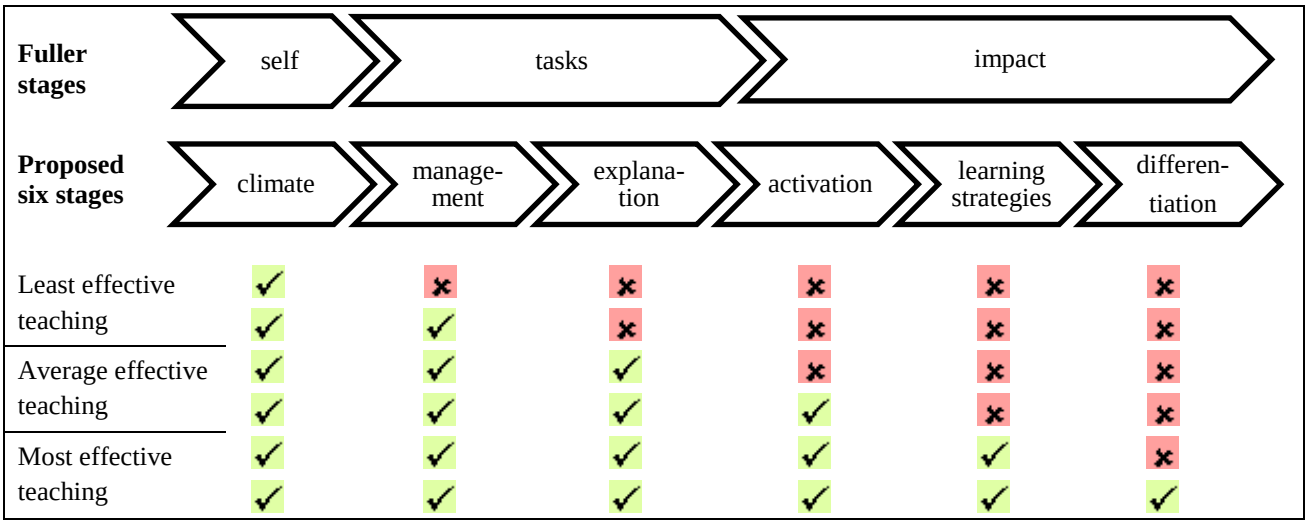


Figure 2.

Hypothetical sample of teachers in which the actual skill level of teachers 1, 2, and 3 is well defined by the stage model, but teacher 4's is not.

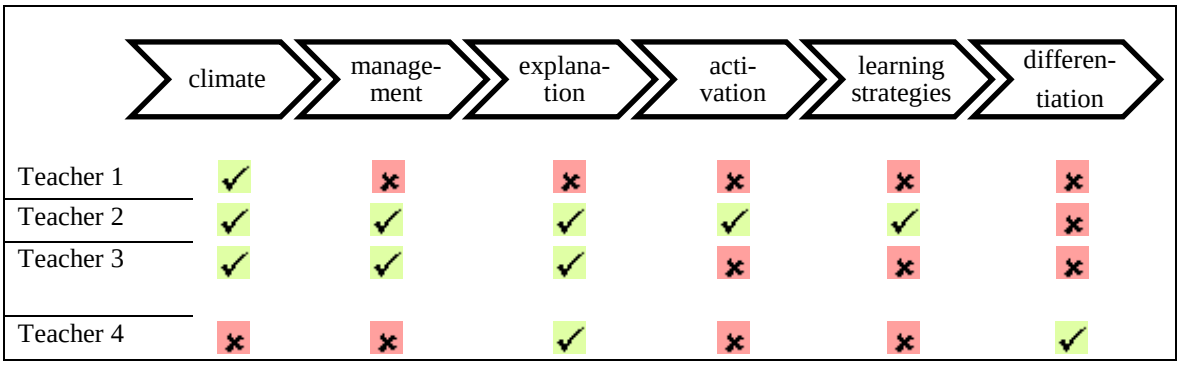
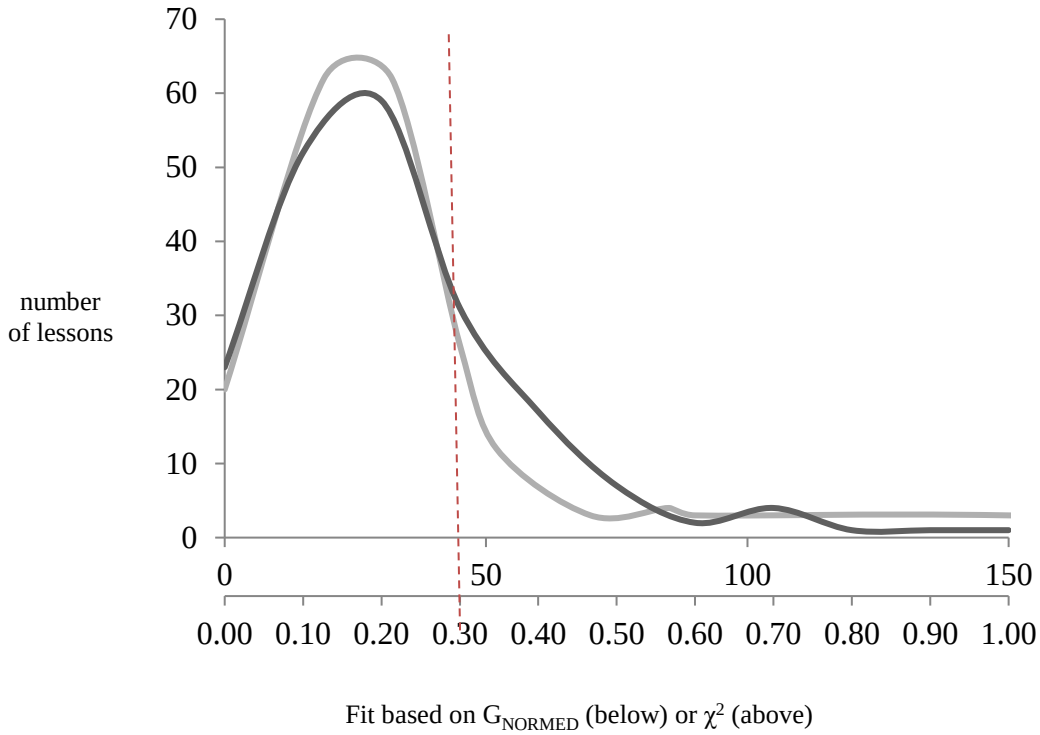


Figure 3.

Graphical representation of G_{NORMED} (black line) and χ^2-person-fit (gray line) statistics. The vertical line accentuates the cutoff values we use for both statistics.



Notes: The gray chi-square distribution refers to the first and upper x-axis, and the black G_{NORMED} distribution refers to the second and lower x-axis. The y-axis shows the number of classroom observations.